Complétion de communautés par l'apprentissage d'une mesure de proximité

Maximilien Danisch, Jean-Loup Guillaume, Bénédicte Le Grand

Complétion de communautés par l'apprentissage d'une mesure de proximité †

Maximilien Danisch[1,2], Jean-Loup Guillaume[1,2] et Bénédicte Le Grand[3]

[1] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6,F-75005, Paris, France*
[2] *CNRS, UMR 7606, LIP6, F-75005, Paris, France*
[3] *CRI, Université Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris, France*

Étant donné un ensemble de nœuds d'un graphe, nous proposons une méthode par apprentissage d'une mesure de proximité afin de compléter, si possible, cet ensemble de nœuds en une communauté. Si cette complétion est possible, la décroissance des proximités a alors une forme plateau/décroissance permettant une détection précise et rapide de la communauté, dite *multi-ego-centrée*. Nos résultats sont validés sur un grand ensemble de pages Wikipédia et sur des benchmarks. L'apprentissage et la détection de plateau/décroissance sont les deux contributions majeures de l'article.

Keywords: détection de communautés, communauté multi-ego-centrée, mesure de proximité, apprentissage

1 Contexte et travaux connexes

L'explosion des réseaux sociaux en ligne a créé d'innombrables opportunités d'établir de nouveaux contacts et de partager des informations, mais elle pose également de nouvelles questions qui constituent un défi de la plus haute importance pour la communauté scientifique : "Avec quel(s) contact(s) Facebook partager un message ?", "Quelle(s) page(s) Wikipédia lire en priorité pour en apprendre le plus possible sur un sujet ?". Deux notions structurelles sont centrales pour appréhender ces problèmes : *les communautés* qui donnent une description de la structure d'un réseau en identifiant des groupes de nœuds fortement liés [For10] et *les mesures de proximités* qui indiquent dans quelle mesure deux nœuds sont topologiquement proches. Ces deux notions, fortement intriquées, constituent le sujet de cet article.

Plus précisément, nous présentons une mesure de proximité paramétrée entre nœuds. À partir d'un ensemble de nœuds donné, nous apprenons les paramètres de la mesure, de sorte que ces nœuds soient proches les uns des autres. Cela nous permet ensuite d'obtenir la proximité de chaque nœud du graphe à l'ensemble donné. Si l'ensemble de nœuds initial appartient à une même communauté, en triant les proximités par ordre décroissant, on observe une courbe formée d'un plateau et suivie d'une forte décroissance qui permet une détection précise et rapide de la communauté, dite *multi-ego-centrée*.

Nous validons ensuite nos résultats sur un réseau de 40 millions de pages Wikipédia connectées par 200 millions d'hyperliens (rendus asymétriques). Pour cela, nous nous intéressons à la catégorie "Graph theory" annotée par les utilisateurs. Cette catégorie comporte 542 nœuds que nous séparons en un training set et un test set de 271 nœuds chacun. Nous validons également nos résultats sur le benchmark de [LF09].

Les travaux les plus connexes sont [KNV06, TF06] où les auteurs cherchent les "k" nœuds les plus pertinents pour un ensemble de nœuds fixé, avec une méthode à base de mesure de proximité ; [SG10] où les auteurs cherchent à compléter un ensemble de nœuds en une communauté en optimisant de manière gloutonne une fonction de qualité (le degré minimum du graphe induit par les nœuds choisis) mesurant combien un ensemble de nœuds donné est une bonne communauté ; [DGLG13] où nous avons montré qu'un petit ensemble de nœuds est généralement suffisant pour caractériser une communauté. L'apprentissage, l'apparition de structure plateau/décroissance en présence de communauté ainsi que la possibilité de caractériser un nœud en fonction de sa centralité vis-à-vis de l'ensemble des nœuds donnés en entrée constituent les contributions majeures de ce papier par rapport à ces travaux.

† Ce travail est partiellement soutenu par l'ANR, projet CODDDE ANR-13-CORD-0017-01. Les auteurs remercient D. Obradovic pour son aide avec les données Wikipédia et S. Kirgizov et D. F. Bernardes pour les conseils et discussions.

2 Méthodologie

Considérons une paire de nœuds n_1 et n_2 connectés mais isolés du reste du réseau et une paire de nœuds n_3 et n_4 non connectés mais partageant N voisins (le module n_3, n_4 et les N voisins sont isolés du reste du réseau). Qui sont les plus proches : (n_1,n_2) ou (n_3,n_4) ? Cette question dépend de N et d'une compréhension du réseau. La fonction de proximité doit donc être contrôlée par un ou plusieurs paramètres. Cette problématique est généralisable aux chemins de longueur supérieure à 2 : deux nœuds non connectés mais avec un voisin en commun sont-ils plus ou moins proches que deux nœuds non connectés, sans voisin en commun, mais ayant N chemins de longueur 3 entre eux ?

Un autre problème vient de la présence de nœuds de fort degré (ou hubs). Soit un nœud n_1 choisi au hasard dans le graphe : est-ce qu'un nœud n_2 fortement lié à n_1 et à son voisinage et faiblement lié au reste du graphe est plus proche de n_1 qu'un nœud n_3 fortement lié partout (y compris au nœud n_1 et à son voisinage) ? Ici aussi le choix doit être laissé libre et nécessite donc de paramétrer la mesure de proximité.

Pour résoudre ces deux problèmes, nous proposons une mesure de proximité qui est une variante de l'indice de Katz. La proximité entre i et j vaut

$$P_{\alpha,\beta,\lambda,\delta}(i,j) = \sum_{l=0}^{\lambda} \gamma_{l,d_j} C_l(i,j) \quad \text{,avec} \quad \gamma_{l,d_j} = \begin{cases} \frac{\alpha^l}{d_j^\beta} & \text{si } d_j \geq \delta \\ \frac{\alpha^l}{\delta^\beta} & \text{si } d_j < \delta \end{cases} \tag{1}$$

où d_j est le degré du nœud j et $C_l(i,j)$ le nombre de chemins de longueur l entre les nœuds i et j. α et λ et permettent de se limiter aux chemins de petite taille ; β et δ permettent d'handicaper les nœuds en fonction de leur degré. Nous expliquerons plus avant ces quatre paramètres par la suite.

Le nombre de chemins de longueur l entre les nœuds i et j étant difficile à calculer (bien que des techniques d'approximation existent), nous proposons d'utiliser le nombre de "non-backtracking paths" (ou NBP), qui sont des chemins n'autorisant que les boucles de taille trois ou plus. Le nombre de NBP d'un nœud i à tous les nœuds se calcule à l'aide du système d'équations suivant, en posant X_l le vecteur contenant le nombre de NBP de longueur l (i.e. la $j^{\text{ème}}$ coordonnée correspond au nombre de NBP entre i et j) :

$$\begin{array}{rclcrcl} X_0 & = & \delta_i & & X_2 & = & AX_1 - DX_0 \\ X_1 & = & AX_0 & \forall l \geq 3, & X_l & = & AX_{l-1} - (D-I)X_{l-2}, \end{array} \tag{2}$$

où δ_i est le vecteur nul sauf pour la coordonnée i qui vaut 1, A est la matrice d'adjacence, D est la matrice diagonale des degrés et I est la matrice identité. Le terme $(D-I)X_{l-2}$ élimine les chemins qui reviennent d'un pas en arrière lors du $l^{\text{ème}}$ pas. La complexité pour calculer le nombre de NBP de longueur l ou moins pour un nœud à tous les autres nœuds du graphe est donc linéaire, en $O(l(n+m))$, où n est le nombre de nœuds et m le nombre de liens. Remarquons qu'après avoir calculé et stocké le nombre de NBP entre le nœud d'intérêt et tous les nœuds du graphe, le calcul de la proximité pour un jeu de paramètres donné se fait en $O(n)$. Une optimisation rapide des paramètres est donc possible.

La figure 1 nous montre que, à degré fixé, deux pages de la catégorie "Graph theory" ont significativement plus de NBP de longueur 1, 2 et 3 que deux pages quelconques. Cette propriété devient fausse pour des NBP de longueur 4 et plus. Seuls les NBP de longueur inférieure ou égale à 3 sont donc, ici, informatifs en termes de communauté. On peut donc fixer le paramètre λ à 3.

Le paramètre δ sert à ne pas trop avantager les nœuds de faible degré. Étant donné un nœud i, un noeud de faible degré peu pertinent pour i (par exemple s'il est lié à des hubs eux-mêmes liés à i) peut en effet avoir une plus grande proximité à i qu'un nœud plus pertinent mais de degré plus fort (handicapé par β). Prendre $\delta = 5$ permet de pallier raisonablement ce problème.

L'apprentissage de (α,β) s'effectue de la façon suivante : étant donné un nœud n_1 dans l'ensemble de nœuds donné en entrée, on calcule la proximité de tous les autres nœuds à n_1, puis on sélectionne le jeu de paramètres qui classe le mieux les nœuds de l'ensemble d'entrée. Ce classement est évalué en fonction de l'AUC [‡] et l'apprentissage est fait par force brute sur l'ensemble $\{0.001^{i/100} \mid i \in [\![0,100]\!]\}$ pour α et $\{0.5+0.005i \mid i \in [\![0,100]\!]\}$ pour β, ensembles obtenus après avoir effectué des tests préliminaires.

‡. L'AUC est une mesure de la précision d'un classifieur très utilisée en apprentissage. Elle est égale à la probabilité de classer un objet étiqueté positif avant un objet étiqueté négatif.

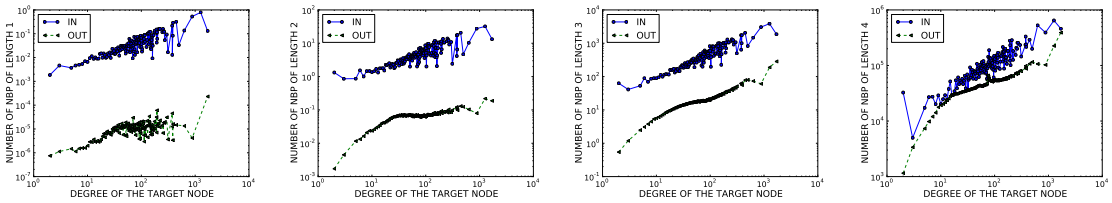


FIGURE 1: Nombre moyen de NBP de longueur 1 à 4 (de gauche à droite) en fonction du degré du nœud d'arrivée pour un nœud source et un nœud d'arrivée dans la catégorie "Graph theory" (IN) d'une part et pour un nœud source dans la catégorie "Graph theory" et un nœud d'arrivée hors de la catégorie "Graph theory" d'autre part (OUT).

La figure 2 montre le classement obtenu à partir de trois nœuds pour les couples optimaux (α, β). Remarquons que la place des nœuds du training set et du test set sont similaires, ce qui montre qu'il n'y a pas d'overfitting. "Global shipping network" donne un mauvais classement (en terme d'AUC, mais également visuellement), car il est périphérique à la communauté "Graph theory", voire en dehors. "Resistance distance" donne un meilleur classement, cependant ce nœud appartient aussi à la communauté des pages parlant d'électricité (il est, entre autres, lié à "Ohm" et "Resistance") ; ce nœud caractérise donc mal la communauté "Graph theory" à lui seul. Le nœud "Multiple edges" enfin donne un très bon classement, il est au centre de la communauté "Graph theory" et suffit donc à lui seul à la caractériser.

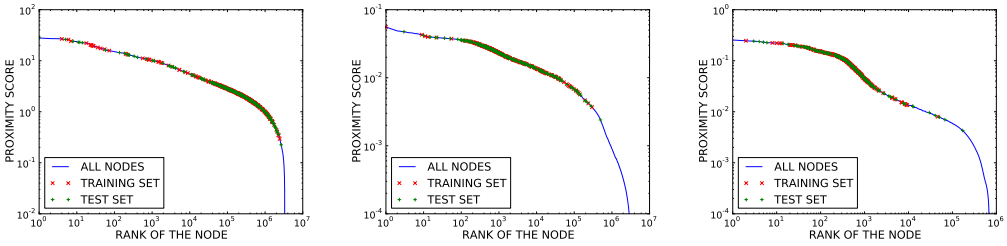
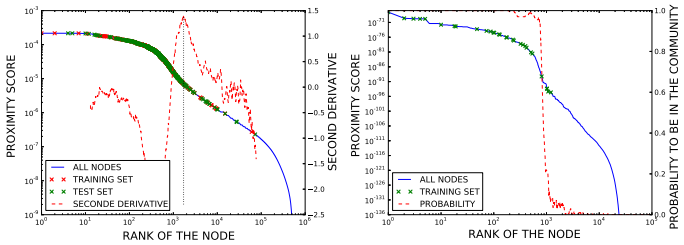


FIGURE 2: De gauche à droite, scores de proximité en fonction du classement pour les couples (α, β) donnant la meilleure AUC en partant des nœuds "Global shipping network", "Resistance distance" et "Multiple edges".

L'AUC du classement obtenu par un nœud permet donc d'évaluer dans quelle mesure le nœud est central pour l'ensemble donné en entrée. Cependant, si l'on considère deux nœuds donnant un classement moyen ("Fan Chung" et "Resistance distance") et que, pour chaque nœud, on effectue le produit des deux scores de proximité on obtient un nouveau classement dont l'AUC est comparable à celle de "Multiple edges" ; cela signifie que l'ensemble {"Fan Chung","Resistance distance"} est suffisant pour caractériser la communauté "Graph theory". Cela montre également que combiner les classements individuels donne de meilleurs classements et on peut donc définir la proximité à un ensemble comme la combinaison (le produit) des proximités individuelles qui donne la meilleur AUC. Dans notre cas il est obtenu par le produit des deuxième et troisième meilleurs classements individuels (voir figure 3(a)).

3 Coupe et validation

La figure 3(a) montre que couper à la plus grande dérivée seconde permet une détection précise de la communauté multi-ego-centrée sur l'ensemble d'entrée. L'étude manuelle des étiquettes des nœuds situés avant et après la coupe montre une transition de pages relatives à "Graph theory" vers des pages ne parlant pas de "Graph theory", y compris une exclusion des nœuds présents à tort dans l'ensemble d'entrée, comme "Global shipping network". L'idée de cette coupe à la dérivée seconde la plus grande est renforcée par des tests effectués sur le benchmark [LF09] (figure 3(b)). D'autres tests suggèrent que toute méthode similaire sans apprentissage, c'est-à-dire utilisant une mesure de proximité sans paramètre, ne peut guère rivaliser.



(a) (b)

FIGURE 3: 3(a) : scores de proximité en fonction du meilleur classement combiné ainsi que la dérivée seconde de cette courbe. 3(b) : classement obtenu pour un graphe avec une structure en communautés recouvrantes généré avec le benchmark [LF09]. Le graphe a 100k nœuds, un degré moyen de 15, 10k nœuds appartenant à 3 communautés (le reste des paramètres est gardé par défaut). La communauté considérée contient 961 nœuds dont 541 appartenant à 3 communautés. L'ensemble input est constitué de 30 nœuds aléatoirement choisis parmi les 961. La courbe probabilité d'être dans la communauté pour un nœud classé k correspond à la proportion de nœuds dans la communauté des nœuds classé entre le $(k-x)^{\text{ème}}$ et le $(k+x)^{\text{ème}}$, pour obtenir une courbe lisse nous avons pris $x = 100$.

4 Conclusion et perspectives

Nous avons présenté une méthodologie qui, étant donné un ensemble de nœuds, calcule un score de proximité entre tous les nœuds du graphe et cet ensemble. Notre méthodologie utilise une proximité paramétrée, apprend ces paramètres et combine les classements individuels obtenus pour chacun des nœuds de l'ensemble donné. L'étude des classements individuels pour un nœud donné permet de savoir s'il est plutôt central ou périphérique dans l'ensemble. Si la majorité des nœuds de l'ensemble de départ appartient à une communauté, une structure en plateau/décroissance de la courbe des scores de proximité (en fonction du classement) est obtenue et une coupe à la dérivée seconde la plus grande permet une détection précise de la communauté, dite multi-ego-centrée. Nous avons validé la méthodologie avec la complétion de catégories (annotées par les utilisateurs) dans Wikipédia et des tests sur des graphes jouets et le benchmark [LF09].

Une possible extension de ces travaux est l'étude de communautés multi-ego-centrées pondérées, possiblement avec des poids négatifs, ainsi qu'une adaptation pour suivre l'évolution des communautés dans les réseaux dynamiques.

Références

[DGLG13] M. Danisch, J.-L. Guillaume, and B. Le Grand. Une approche à base de similarité pour la détection de communautés egocentrées. *15èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel)*, pages 1–4, 2013.

[For10] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3) :75–174, 2010.

[KNV06] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–255. ACM, 2006.

[LF09] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1) :016118, 2009.

[SG10] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 939–948. ACM, 2010.

[TF06] H. Tong and C. Faloutsos. Center-piece subgraphs : problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 404–413. ACM, 2006.